The Dangers of Extreme Counterfactuals

Advance Access publication November 2, 2005

Political Analysis (2006) 14:131–159
doi:10.1093/pan/mpj004

The Dangers of Extreme Counterfactuals

Gary King
Department of Government and Institute for Quantitative Social Science, Harvard University, 1737 Cambridge Street, Cambridge MA 02138
e-mail: king@harvard.edu

Langche Zeng
Department of Political Science, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0521
e-mail: zeng@ucsd.edu

We address the problem that occurs when inferences about counterfactuals—predictions, "what-if" questions, and causal effects—are attempted far from the available data. The danger of these extreme counterfactuals is that substantive conclusions drawn from statistical models that fit the data well turn out to be based largely on speculation hidden in convenient modeling assumptions that few would be willing to defend. Yet existing statistical strategies provide few reliable means of identifying extreme counterfactuals. We offer a proof that inferences farther from the data allow more model dependence and then develop easy-to-apply methods to evaluate how model dependent our answers would be to specified counterfactuals. These methods require neither sensitivity testing over specified classes of models nor evaluating any specific modeling assumptions. If an analysis fails the simple tests we offer, then we know that substantive results are sensitive to at least some modeling choices that are not based on empirical evidence. Free software that accompanies this article implements all the methods developed.

1 Introduction

As recently as a half decade ago, most quantitative political scientists were still presenting statistical results in tables of hard-to-decipher coefficients from logit, probit, event count, duration, and other analyses. These and other models are still in widespread use, but most authors now also compute and present quantities of genuine interest from these models, such as predicted values, expected counts, first differences, causal effects, risk ratios, etc.

Authors' note: Easy-to-use software to implement the methods introduced here, called "WhatIf: Software for Evaluating Counterfactuals," is available at http://GKing.Harvard.Edu/whatif. At the encouragement of the editors, we wrote a companion piece that overlaps this article; it excludes the mathematical proofs and other technical material and has less general notation, but it includes additional examples and more pedagogically oriented material. See "When Can History Be Our Guide? The Pitfalls of Counterfactual Inference," *International Studies Quarterly*, forthcoming (available at http://GKing.Harvard.edu/files/abs/counterf-abs.shtml). Thanks to Jim Alt, Scott Ashworth, Neal Beck, Jack Goldstone, Craig Gotsman, Sander Greenland, Kosuke Imai, Orit Kedar, Walter Mebane, Joe Mitchell, Maurizio Pisati, Kevin Quinn, Jas Sekhon, and Simon Jackman for helpful discussions, and to the National Institutes of Aging (P01AG17625-01), the National Science Foundation (SES-0318275, IIS-9874747), and the Weatherhead Initiative for research support.

Whether such effects are calculated via analytical derivation or what is now the more common approach of statistical simulation (King et al. 2000), political scientists have made much progress in learning how to make sophisticated methods speak directly to their substantive research questions. Although this represents considerable progress in reducing the barriers between technique and substance, we address here a crucial remaining disconnect, one that threatens to undermine the validity of a considerable body of important research. This is the problem of extreme counterfactuals—predictions, what-if questions, and causal inferences that are so far from the data that inferences wind up being drawn on the basis of minor model specification choices no one would like to defend, rather than empirical evidence. The result in these situations is such a high degree of model dependence, even among models that fit well, that, unbeknownst to the authors or readers, analyses can turn out to be theoretical exercises masquerading as empirical estimation.

For example, with a sample of U.S. time series data, we could reasonably ask how much U.S. presidential approval would drop if inflation increased by 2 percentage points, and we could generate a fairly certain answer. However, to take an absurdly extreme alternative for the sake of clarity, we should not expect to get a precise empirically based answer from the same data, given *any* model, if we asked how much approval would drop if inflation increased by 200 percentage points. Any good data analyst would know not to ask this question of the available U.S. data, but exactly what is the principle underlying this decision and how can we learn how to apply the principle in cases where the counterfactual is not so absurdly far from the data? Is it reasonable to ask of the same data what would happen to approval if inflation today increased to 10%? 20%? 30%? Where is the cutoff?

The problem with extreme counterfactuals is that whatever statistical model we used to compute the 2% counterfactual inference could also be used to compute the 200% one. Our confidence interval for counterfactuals farther from the data are wider, but the inference may be considerably more uncertain than the confidence interval indicates. The confidence interval is not wrong: if the other assumptions of the model are correct, it accurately portrays the uncertainties, conditional on the model. The problem is that we have little reason to assume the model is right when the counterfactual is so far from the data. In other words, the 200% inference is far more *model dependent* than the first. Figure 1 illustrates this, where two alternative models, a linear and a quadratic model, are not distinguishable within the range of the data, but the quadratic falls far outside the confidence intervals of the linear model for counterfactuals (i.e., values of X) farther from the data. Thus, if the true data-generating process at a location far from the observed data is actually quadratic, the linear model's prediction would be far from the truth, and even the wide confidence interval would not contain the true value.

The key question, however, is how to tell how model dependent inferences are when the counterfactual is not so obviously extreme or when it involves more than one explanatory variable. Extreme counterfactuals are not always easy to spot, especially given the relatively few quantitative approaches to this problem. The answer to this question does not come from the model-based quantities we normally compute, such as standard errors, confidence intervals, coefficients, likelihood ratios, predicted values, test statistics, first differences, p values, etc. To understand how far from the facts are our counterfactual inferences, and thus how model dependent are our inferences, we need to look elsewhere. At present, scholars study model dependence primarily via sensitivity analyses: changing the model and assessing how much conclusions change. If the changes are substantively large for models in a particular class, then inferences are deemed model dependent and thus unreliable. This is a fine approach, but it is insufficient in circumstances where the class of possible models cannot be easily formalized and identified or where the models

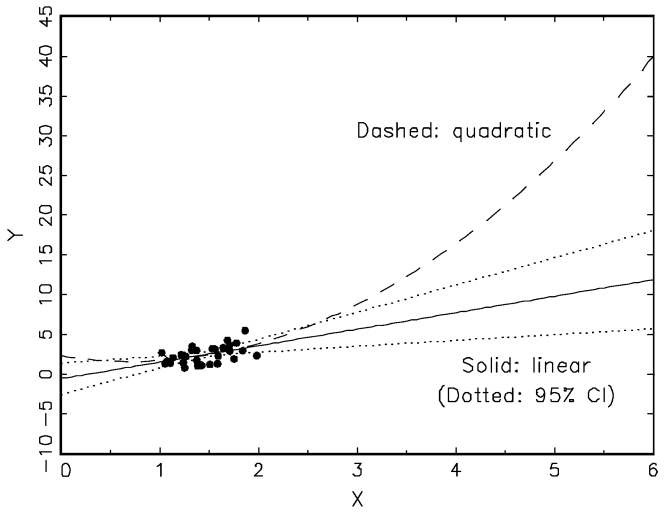

Fig. 1 Model dependence with good in-sample fit.

within a particular class cannot feasibly be enumerated and run, i.e., most of the time. In practice, the class of models chosen are those that are convenient—such as those with different control variables under the same functional form. Moreover, the identified class of models normally excludes at least some models that have a reasonable probability of returning different substantive conclusions. Most often, this approach is skipped entirely.

We provide several easy-to-apply methods that reveal the degree of model dependence, without having to run all the models. The methods apply to the class of nearly all models, whether or not they are formalized, enumerated, and run, and for the class of all possible dependent variables, conditional only on the choice of a set of explanatory variables. If an analysis fails our tests then we know it will fail a sensitivity test too, but without being in the impossible position of having to run all possible models to find out.

Section 2 shows more specifically how to identify questions about the future and "what-if" scenarios that cannot be answered well in given data sets. This section introduces several approaches for assessing how based in factual evidence is a given counterfactual. It also proves formally, apparently for the first time, that inferences about counterfactuals farther from the data are more model dependent. In addition, via a solution to a difficult problem in computational geometry, we are able, also apparently for the first time, to make practical the use of what was a computationally infeasible but conceptually intuitive and widely recognized criterion for identifying counterfactuals that are "too far" from the data.

Section 3 discusses the connection between using extreme counterfactuals in prediction and what-if questions, and that in causal inferences. It also provides a new decomposition of the bias in estimating causal effects using observational data more suited to the problems most prevalent in political science than the best available decomposition in the literature. This decomposition enables us to identify causal questions without good causal answers in given data sets and shows how to narrow these questions in some cases to those that can be answered more decisively. We also show how the new results introduced in section 2 are useful for identifying which counterfactuals in causal inference will be model dependent and thus not sufficiently based on the evidence to draw reasonably valid conclusions, something to which existing approaches, most based on the propensity score, are not well suited.

2 Forecasts and "What-If" Questions

Although statistical technology sometimes differs for making forecasts and estimating the answers to what-if questions (e.g., Gelman and King 1994), the logic is sufficiently similar that we consider them together here. In regression-type models, including least squares, logit, probit, event counts, duration models, and most others used in the social sciences, we usually compute forecasts and answers to what-if questions using the model-based conditional expected value of Y, the dependent variable, given a chosen vector of values x of the explanatory variables, X.[1] (Thus x is the same dimensions as a row of X.) The model typically includes a specification for (i.e., assumption about) the conditional expectation function (CEF):

$$\mathrm{E}(Y \mid X) = g(X, \beta), \tag{1}$$

where $g(\cdot)$ is some specified parametric functional form and β is a vector of parameters to be estimated. To make a forecast, we plug the specified vector of values x, and the point estimate of β, which we denote $\hat{\beta}$, into this CEF and compute the estimated CEF or predicted value:

$$\hat{\mathrm{E}}(Y \mid x) = g(x, \hat{\beta}). \tag{2}$$

The estimated CEF for the familiar linear regression model, for example, is $x\hat{\beta} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$; the logit model is $[1 + e^{-x\hat{\beta}}]^{-1}$; the exponential duration model is $e^{-x\hat{\beta}}$; count models are usually specified as $e^{x\hat{\beta}}$, etc.[2]

Interestingly, each of these CEFs can be computed for *any* (real) values of x. The model never complains, and exactly the same calculation can be applied for any value of x. However, even if the model fits the data we have in our sample well, a vector x far from any rows in the matrix X is not likely to produce accurate forecasts. If a linear model indicates that one more year of education will earn you an extra $1,000 in annual income, the model also implies that 10 more years of education will get you $10,000 in extra annual income. In fact, it also says that 50 years more of education will raise your salary by $50,000. At 50 years of education, the counterfactual is so far from the data that it is downright silly. But somewhere past one year, but well before the question becomes obviously silly, comes a distance from the data at which inferences become sufficiently model dependent that conclusions become based more on small modeling assumptions than on the data. This article is devoted to understanding this gradation. The key is that even though no statistical assumption may be violated as a result of the choice of any set of real numbers for x, the model obviously produces better forecasts (and what-if evaluations) for some values of x than others, but no measure produced by standard statistics packages helps guide research in choosing reasonable counterfactuals. Predictive confidence intervals for forecasts farther from the data are usually larger, but confidence intervals computed in the usual way still assume the veracity of the model no matter how far the counterfactual is from the data.

[1] For notational convenience we also use Y and X to denote the observed data matrix of the dependent and explanatory variables when the context is clear.

[2] More generally, we are interested in the full conditional density, $\mathrm{P}(\mathrm{Y} \mid \mathrm{x}) = \int \mathrm{P}(\mathrm{Y} \mid \mathrm{x}, \beta)\mathrm{P}(\beta \mid \mathrm{Y})\mathrm{d}\beta$. All our methods apply in this situation as well, but for expository purposes we continue to focus on the CEF in Eq. (1). Eq. (2) is included only to fix ideas, since our methods do not require an estimate of the CEF or a specification of the model. Normally a better way of computing the estimated CEF recognizes the uncertainty in $\hat{\beta}$: $\hat{\mathrm{E}}(\mathrm{Y} \mid \mathrm{x}) = \int \mathrm{g}(\mathrm{x}, \hat{\beta})\mathrm{P}(\hat{\beta})\mathrm{d}\hat{\beta}$, where $\mathrm{P}(\hat{\beta})$ is the posterior density of β.

A key point is that more careful model choice will not help here, since far away from the observed data we simply have no empirical evidence to test or with which to compare models. Other models will not do verifiably better with the same data, and evaluating the evidence to see which model, among those that fit, is "better" cannot help with a counterfactual not near the data being used for evaluation. So searching for a better model, without better data, better theory, or a different counterfactual question, in this case is simply futile. We merely need to recognize that some questions cannot be answered from some data sets. Our linearity (or other functional form assumptions) are written globally—for any value of x—but in fact are verifiable only locally—in or near our observed data. In this article, we seek to provide some tools to help ascertain where "local" ends and "global" begins. For forecasting and analyzing what-if questions, our task comes down to seeing how "far" the point x is from the observed data matrix X.

2.1 *Model Dependence as a Function of Distance from the Data*

Before turning to the tools, we prove formally in this section that counterfactual inferences farther from the data are more model dependent. In our proof we do not assume knowledge of specific functional forms, models, estimators, or dependent variables, or a specific definition of "distance" from the data. We do make assumptions, but it turns out that much less restrictive assumptions are sufficient.

To put model-based inferences in context, consider first a model-free inference. When many rows of X contain replicas of the exact posited counterfactual question, x, model-free inference is possible. To estimate the CEF $\mathrm{E}(Y \mid X = x)$ at the point x, in this situation, we simply take the average of the observed Y values among observations for which $X = x$. If other standard assumptions are met, then no assumption about the functional form is required and model dependence is not a risk.

In most situations, however, the data X contain no values (or too few values) that correspond exactly to the counterfactual x. Finding a country just like the United States in all measured respects except with very low GDP or observing presidential approval for a year that has not yet occurred are simple examples of such counterfactuals. In these situations, inference is made possible via modeling assumptions, so model dependence then becomes a risk. Indeed, if we make no assumptions at all about the true CEF, then except in special cases learning about counterfactuals other than those that are coincident with the data points X (the "factuals") is impossible.

For simplicity in this section, we define *model dependence* at point x as the difference, or distance, between the predicted outcome values from any two *plausible* alternative models. (One model might be logit and the other probit, or one linear the other quadratic, etc.) By "plausible" alternative models, we mean models that fit the data reasonably well and, in particular, they fit about equally well around either the "center" of the data (such as a multivariate mean or median) or the center of a sufficiently large cluster of data nearest the counterfactual x of interest. It is easy to generate model dependence when the model does not even fit the data, but this is easy to avoid, as many current data analysis techniques are designed to detect and correct poorly fitting models.

Predictions are typically obtained by evaluating the conditional expectation functions of the two models, $g_1(x)$ and $g_2(x)$, at x. Model dependence at point x is thus the distance from $g_1(x)$ to $g_2(x)$, or the *norm* $\|g_1(x) - g_2(x)\|$ (which is the distance from the vector to 0). We do not restrict the types of distance measures in the proof. It could be defined simply as the absolute or Euclidean distance between the two, but any other distance measures suitable for a given application would work too. In all cases, of course, the distance

induced by the normed space is greater than or equal to zero, is exactly zero only when $g_1(x) = g_2(x)$, and has the other mathematical properties that qualify it as a proper distance (including symmetry and the triangle inequality).

Denoting X^* as a location in the data that represents the center of the entire data set or a sufficiently large portion of the data near our counterfactual, our assumption is that the two functions evaluated at this point give about the same value for the CEF:

$$\|g_1(X^*) - g_2(X^*)\| \approx 0. \tag{3}$$

Our assumption is not restrictive. We do not assume that either of the functional forms fit well by any absolute standard, but only that neither fits much better than the other where data are plenty.

To prove that model dependence is a function of distance from the data (in the specific sense that its lowest, or "sharpest," upper bound available is a function of the distance from the data), we make only one other assumption, that the conditional expectation functions of alternative models behave reasonably well in the sense of satisfying a strong continuity condition named the Lipschitz condition, on a convex set containing both the observed data X and the counterfactual x, so that for any two points on this set, and in particular x and X^*,

$$\|g_1(x) - g_1(X^*)\| \leq L_1\|x - X^*\| \tag{4}$$

$$\|g_2(x) - g_2(X^*)\| \leq L_2\|x - X^*\| \tag{5}$$

where L_1 and L_2 are finite positive constants. This is a quite weak assumption, which requires continuity plus bounded slopes. Thus it rules out discontinuous functions such as step functions (but not when the discontinuity is coded in the explanatory variables) and functions that veer off to infinity between two points a finite difference apart, but it easily includes the vast majority of the functional forms regularly specified in the journals throughout the social sciences. Having bounded derivatives is sufficient but not necessary for Lipschitz continuity to hold, since Lipschitz continuous functions need not be differentiable.

One way to understand the Lipschitz condition is in the case in which X contains only a single column. In this case, the condition can be written as

$$\left|\frac{g_1(x) - g_1(X^*)}{x - X^*}\right| \leq L_1$$

and similarly for g_2. This expression means that the slope of the line joining any two points on the graph (in particular x and X^*) is bounded.

Under these assumptions, we derive an expression for the degree of model dependence at counterfactual point x as a function of the distance from x to the data X^*:

$$\|g_1(x) - g_2(x)\| = \|[g_1(x) - g_1(X^*)] - [g_2(x) - g_2(X^*)] + [g_1(X^*) - g_2(X^*)]\| \tag{6}$$

$$\leq \|g_1(x) - g_1(X^*)\| + \|g_2(x) - g_2(X^*)\| + \|g_1(X^*) - g_2(X^*)\| \tag{7}$$

$$\leq (L_1 + L_2)\|x - X^*\| + \|g_1(X^*) - g_2(X^*)\| \tag{8}$$

$$\approx (L_1 + L_2)\|x - X^*\| \tag{9}$$

where Eq. (6) holds by adding and subtracting $g_1(X^*)$ and $g_2(X^*)$, Eq. (7) holds by the properties of norms, Eq. (8) substitutes in the definition in Eqs. (4) and collects terms, and Eq. (9) applies Eq. (3).[3]

Equation (9) is highly informative. It shows that for two models that fit about as well as each other, the maximum degree of model dependence is in effect solely a function of the distance from the counterfactual point to the data. Thus the farther the counterfactual point is from the data, defined as X^*, the more likely our inferences are to be model dependent. Finally, we note that this is a quite general result since it holds for *any* two alternative models whose conditional expectation functions g_1 and g_2 satisfy the Lipschitz condition and fit the data approximately as well, not only those we might think to check.

2.2 *Measuring Distance from the Data*

We now offer tools that measure how "far" a counterfactual is from the data. We begin with a general distance measure in this section. Section 2.3 then simplifies by introducing a dichotomous criterion. Both are useful in practice with variables of any type.

Our goal here is some measure of the fraction of observations (rows) in X "near" the counterfactual x. To create this measure, we begin with a measure of distance between two points (or rows) x_i and x_j based on Gower's (1966, 1971) measure. This is only one possible choice, but it is a reasonable one that applies to most data types, including discrete and continuous variables. It is defined simply as the average absolute distance between the elements of the two points, divided by the range of the data:

$$G_{ij}^2 = \frac{1}{K}\sum_{k=1}^{K}\frac{|x_{ik} - x_{jk}|}{r_k}, \tag{10}$$

where the range is $r_k = \max(X_{.k}) - \min(X_{.k})$ and the min and max functions return the smallest and largest elements, respectively, in the set including the kth element of the explanatory variables X. Thus the elements of the measure are normalized for each variable to range between zero and one, and then averaged. The measure is designed to apply to all types of variables, including both continuous and discrete data.[4] If $G^2 = 0$, then x and the row in question of X are identical, and the larger G_{ij}^2, the more different the two rows are. We interpret G^2 as *the distance between the two points as a proportion of the distance across the data*, X.[5] So $G^2 = 0.3$ means that to get from one point to the other, one needs to travel the equivalent distance as 30% of the way across the data set.

With G^2 applied to our problem, we need to summarize n numbers, the distances between x and *each* row in X. If space permits, we suggest presenting a cumulative frequency plot of G^2 portraying the fraction of rows in X with G^2 values less than the given value on the horizontal axis. If space is short, such as would happen if many counterfactuals need

[3]If assumptions can be added in an application to narrow the range of functions allowed further than the Lipschitz condition, then model dependence can be shown to depend even more strongly or in specific ways on the distance between the counterfactual and observed data (see, for example, Madych and Nelson 1992; Wu and Schaback 1993; and Shaback 1996). The same qualitative conclusion would, of course, still hold.

[4]Ordinal explanatory variables are typically assumed interval or coded as a set of dichotomous variables, and Gower's measure follows that practice. Nominal multichotomous variables are also coded as a set of dichotomous variables. Some versions also include weights, but we exclude those here. We recommend choosing r_k from the sampling design if X was selected by stratification or experimental manipulation, or as above if X was randomly sampled. Gower (1971) shows that G satisfies the triangle inequality, so his measure and our simple modifications of it therefore have metric interpretations.

[5]Technically speaking, G is the measure shown to satisfy the mathematical properties of a "distance," but we use the word qualitatively to apply to G^2, which has a more intuitive substantive interpretation.

to be evaluated, any fixed point on this graph could be used as a one-number summary. Our recommendation for a rule of thumb in defining observations that are sufficiently close to the counterfactual to make for reasonable inferences is to use the fraction (or number) of observations in the data with distances (values of G^2) less than the "geometric variability" (GV) of X—which is roughly the average distance among all pairs of observations in the data.[6] We interpret the resulting measure—based on the observations less than one GV away from the counterfactual—as the fraction of the observed data *near* the counterfactual. We have found this rule of thumb to be useful in practice for determining the effective number of observations available to make inferences without high levels of model dependence.

Observations farther than one GV away from the counterfactual normally have little empirical content for inference about the counterfactual and can produce considerable model dependence. Researchers should consider downweighting or even discarding these observations from the data, unless they are in the unusual situation of being certain that their model specification is correct.[7]

2.3 *Interpolation versus Extrapolation*

We now offer a simpler summary measure of how far the counterfactual is from the data. This is the distinction between whether computing the counterfactual $\mathrm{E}(Y \mid x)$ would involve *interpolation* or *extrapolation*. Perhaps the most intuitive definition of extrapolation is one based on the concept of the *convex hull*: questions that involve interpolation are values of the vector x that fall in the convex hull of X, and those involving extrapolation are outside of the hull. This distinction is well known and accepted in the statistical literature, but it has not been used in real applications with more than an explanatory variable or two (e.g., Kuo 2001; Hastie et al. 2001) due to computational difficulties in finding the convex hull for high-dimensional data and determining whether points fall in it. Below we first explain the intuitive meaning of the convex hull criterion, then discuss the computational issues and a solution.

2.3.1 Convex Hull

The convex hull for one variable is bounded by the maximum and minimum data points: any counterfactual question between those points requires interpolation; points outside involve extrapolation. For two explanatory variables, the convex hull is given by a polygon with extreme data points as vertices such that for any two points in the polygon, all points that are on the line segment connecting them are also in the polygon (i.e., the polygon is a convex set). This is easiest to see graphically, such as via the example in Fig. 2, given simulated data. A counterfactual question x that appears outside the polygon requires extrapolation. Anything inside involves interpolation.

Although Fig. 2 only portrays the convex hull for two explanatory variables, the concept is well defined for any number of dimensions. For three explanatory variables, and

[6] The geometric variability is also known as the generalized variance (Cuadras and Fortiana 1995; Cuadras et al. 1997) and is what we would refer to as the squared generalized standard deviation. It is a generalized version of the usual variance definition in that for Euclidean distances (which are inappropriate with binary data, for example), it equals the regular variance for one explanatory variable, or in general the trace of the covariance matrix of X. For other measures of distance, such as used in Eq. (10), the geometric variability is a generalized measure of dispersion of X.

[7] Of course, this is only a rule of thumb and so more data-conserving rules could be applied (such as discarding data only 1.5 or 2 GVs away from the counterfactual), as could rules that tolerate even less model dependence, depending on how much confidence one puts in the chosen model. The advice in this paragraph to discard data that make inferences model dependent applies only in the usual situation in which the model is not known. In these situations, the estimator will not normally be "self-efficient" and so the usual "more data are better" rule does not apply. See Meng and Romero (2003).

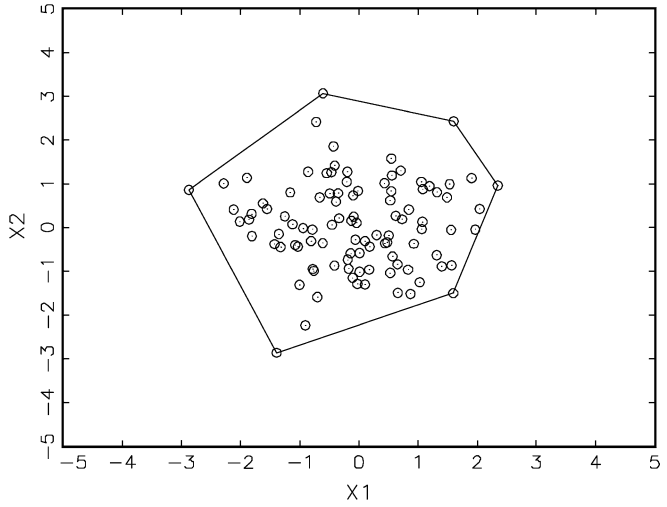

Fig. 2 Interpolation vs. extrapolation: The convex hull of X is the smallest convex set that contains the data. Inference on points inside the convex hull requires interpolation; inference outside it requires extrapolation.

thus three dimensions, the convex hull could be found by "shrink wrapping" the fixed points in three-dimensional space (e.g., see the animation at http://gking.harvard.edu/whatif). The shrink-wrapped surface encloses counterfactual questions requiring interpolation; those falling outside require extrapolation. For four or more explanatory variables, the convex hull is more difficult to visualize, but the mathematical conceptualization is straightforward. The general mathematical definition of the convex hull of a set of points is the smallest convex set that contains them.[8]

Intuitively, counterfactuals outside the convex hull of the observed data are generally farther away from the data. Thus, by the proof in section 2.1, answering a question involving extrapolation is generally more model dependent and thus more hazardous than one involving interpolation. If we learn that a counterfactual question involves extrapolation, we still might wish to proceed if the question is sufficiently important, but we would be aware of how much more model dependent our answers would be. With this definition, and the software we offer, researchers can easily check whether a particular question requires extrapolation.[9]

[8] A set is convex if, for any two elements in the set, the convex combinations of them are also in the set. A point is a convex combination of two other points if it lies on the line segment between the two points, i.e., it is a linear combination of the two points with coefficients that are each between zero and one and together sum to one. See Valentine (1964, p. 13ff).

[9] An alternative formal definition of extrapolation is an inference that occurs at x that is off the *support* of X (that is, the values of the X that have nonzero density), so that there is zero probability of having observations within some neighborhood of x in repeated sampling. Manski (1995, p. 16) uses this definition and shows that for continuous functions, interpolation enables nonparametric identification of the conditional distribution $P(Y \mid X = x)$, (and therefore the CEF at x), while extrapolation requires additional assumptions. Unfortunately, estimating the support of X from sample data is difficult or infeasible for more than a few explanatory variables (and is not required for regression-type models that condition on X), so in this article we focus on our definition using the convex hull, which leads to easy verification. For finite samples, the convex hull of X and the support of X are closely but qualitatively related; there is no universal quantitative relationship. As more and more observations on X are drawn from the same population, the convex hull either equals the support of X or contains it as a subset and so can be seen as a conservative approach.

Since the Gower distance described in section 2.2 is measured as a proportion of the distance across the data, if the Gower distance between x and any row in X is greater than 1, then x lies outside the convex hull of X. The reverse does not necessarily hold. Although counterfactuals outside of the hull are generally farther away from the data, there can be exceptions, as in the case in which most data lie close to the boundary of the hull and the counterfactual lies in a void in the middle. For this reason the distance measure is a useful complement of the convex hull criterion.

Before turning to computational issues, we note that using extrapolation to detect counterfactuals too far from the data is conditional on a specific choice for X, just as is the case in most social science regression-type models. Thus we need make no special accommodations for nonlinearities or interactions. Similarly, we assume outliers are removed as part of the important data preprocessing procedures normally used in standard statistical modeling.[10]

2.3.2 Computational Geometry Problems and Solutions

Unfortunately, there is a serious computational problem with our plan to use convex hull membership as a way to evaluate counterfactuals in practice: Algorithms to identify the convex hull from a set of input data X are enormously time consuming even on very fast computers. The core problem is that the number of facets (or "hyperfacets") a high-dimensional convex hull can have is on the order of $n^{k/2}$ for k variables and n observations (Klee 1980), so even listing all the facets would be impractical. Problems with more than eight or nine variables appear to have been attempted only rarely, if ever. Of course, restricting ourselves in this way would rule out numerous social science analyses, which often have many more covariates.

Moreover, once we have identified the hull we still need to determine whether the counterfactual point x falls within it. In computational geometry, this problem is known as "point location," and in two dimensions it is equivalent to an algorithm that takes the latitude and longitude coordinates of a point on a map and returns the country in which this point falls. Unfortunately, "in more than two dimensions, the point location problem is still essentially open" (de Berg et al. 1998 p. 144).

Perhaps the enormous computational complexity of finding the hull and the problem of point location in higher dimensions accounts for why, although many scholars in the statistical community talk about using the convex hull to define extrapolation, we have found no research that uses it for the kinds of high-dimensional practical problems that commonly occur in social science research. Moreover, no statistical software we have

[10] In the inadvisable situation in which a researcher ignores the problem and persists with checking whether the counterfactual is outside the convex hull, outliers in X would make this extrapolation-detection method overly conservative in identifying counterfactuals that require extrapolation. That is, some counterfactuals x would be identified as requiring interpolation even though they would really involve extrapolation. However, if this method identified a counterfactual as requiring extrapolation, then the suspicion of outliers in X would only make the finding more solid. Similarly, if specific forms of nonlinearity and interactions (such as squared terms or products of existing variables) are known to be present in the CEF, explicitly using them as additional input variables will result in a CEF that is a smoother function of the larger set of inputs. As a result, counterfactual inference within the convex hull of this larger data matrix X will be more accurate, so the results of the test we propose will be more informative about the approximation error in making interpolations. Identification of "features" of the original data such as squared terms or interactions that may be present in the CEF is part of the routine data 'pre-processing' step that political scientists often perform. For a rigorous treatment of this topic see, for example, Bishop (1995, chap. 8). However, researchers should not include these extra terms in their input data X unless they know they belong in the CEF; putting them in when they do not belong could cause one to conclude incorrectly that a counterfactual requires extrapolation.

found includes a procedure to ascertain whether a counterfactual is outside a high-dimensional convex hull and therefore requiring extrapolation.

To make the convex hull criterion of use in practical research, we derived our own solution to the convex hull membership check problem (see Appendix A). The result with even very large numbers of variables is an algorithm that finishes in seconds. The key to our approach is a way to determine whether the counterfactual x falls within the convex hull of of the data X without ever characterizing the convex hull itself. This strategy thus eliminates the most time-consuming part of the problem. In addition, we show how the remaining (implicit) point location problem can be expressed as a linear programming exercise, making it possible to take advantage of existing well-developed algorithms designed for other purposes to speed up the result. We also offer easy-to-use software, "WhatIf: Software for Evaluating Counterfactuals," that automates our algorithm and implements the other methods discussed in this article (see Stoll et al. 2005).[11]

2.4 *Democracy Counterfactuals*

We now apply these methods of evaluating counterfactuals to address one of the most asked questions in political science: what is the effect of a democratic form of government (as compared to less democratic forms)? We study counterfactuals relating to the degree of democracy using data collected by the State Failure Task Force (Esty et al. 1998). See King and Zeng (2002) for an independent evaluation.

This dataset is among the best ever collected in this area. The task force's dependent variable is the onset of state failure, but our analyses apply to all dependent variables. "What would happen if more of the world were democratic" is a question that underlies much other work in comparative politics and international relations over the last half century as well as a good deal of American foreign policy. After extensive searches Esty et al. (1998) predicted state failure with trade openness (as a proxy for economic conditions and government effectiveness), the infant mortality rate, and democracy. Democracy is coded as two dummy variables representing autocracy, partial democracy, and full democracy. King and Zeng (2002) added to these the fraction of the population in the military, population density, and legislative effectiveness.

To see how widely our analyses apply, we began collecting other articles in the field that use a set of explanatory variables with a fair degree of overlap with the set used here, and stopped at 20 after searching only the last few years. The methods presented in this section would need to be repeated to draw more certain conclusions from each of these other articles, but the overlap was sufficient to infer that the results presented here will likely apply without modification to a large number of articles in our field.

We begin our empirical analyses with four clear examples, the first two obviously extrapolations and the second two obviously interpolations, and then move to averages of many other cases of more substantive interest. Before turning to empirically reasonable counterfactuals, we begin with examples that are deliberately extreme. Extreme examples are of course useful for ensuring expository clarity, but they are also useful here since,

[11] After developing our approach and distributing our paper, we learned that there exists an obscure note by Kallay (1986) that offers a related alternative solution to this problem (our thanks to Joe Mitchell for pointing this out). Kallay's article deserves to be better known; it does not appear in textbook reviews of computational geometry ((de Berg et al. 1998; O'Rourke 1998)), has only once in the 20 years since its publication been cited in a research article (according to Google Scholar, accessed on September 4, 2005), and apparently is completely unknown to the statistics community. Our solution is simpler than Kallay's, formulates the problem in terms closer to the statistical issue we are studying, and, since we were able to express the linear programming problem with a degenerate objective function, should usually be about twice as fast on average.

although almost no serious researcher would expect the data to provide information about such counterfactuals if intentionally asked, almost all empirical analysts estimating the effects of democracy have implicitly asked precisely these questions. This is always the case when all observations are used in the estimation and causal effect evaluation, as is typical in the literature. So although the two examples we now introduce are obviously extreme, we show that many questions actually asked in the literature are in fact also quite extreme.

We begin by asking what would have happened if Canada in 1996 had become an autocracy, but its values on other variables had remained at their actual values. We find, as we would expect, that this extreme counterfactual is outside the convex hull of the observed data and therefore requires extrapolation. In other words, we can ask what would have happened if Canada had become autocratic in 1996, but we cannot use history as our guide, since the world (and therefore our data) includes no examples of autocracies that are similar enough to Canada on other measured characteristics. Similarly, if we ask what would have happened if Saudi Arabia in 1996 had become a full democracy, we would also be required to make an extrapolation, since it too falls outside the convex hull.

We now ask two counterfactual questions that are as obviously reasonable as the last two were unreasonable. We ask what would have happened if Poland had become an autocracy in 1990 (i.e., just after it became a democracy). From qualitative information available about Poland, this counterfactual is quite plausible, and many even thought (and worried) about it actually occurring at the time. Our analysis confirms the plausibility of this suspicion since this question falls within the convex hull; analyzing it would require interpolation and thus not much model dependence. In other words, the world has included examples of autocracies that are like Poland in other respects, so history can be our guide. Another reasonable counterfactual is to ask what would have happened had Hungary become a full democracy in 1989 (i.e., just before it actually did become a democracy). This question is also in the convex hull and would require only interpolation to produce specific estimates.

We now further analyze these four counterfactuals using our modified Gower distance measure. The question is how far the counterfactual x is from each row in the observed data set X, so the distance measure applied to the entire data set gives n numbers. We summarize these numbers, without loss of information, in the cumulative frequency plots in Fig. 3. The left plot includes counterfactuals that change from autocracies to democracies, and the right plot is the reverse. The dark line in each graph refers to a counterfactual within the convex hull and the dashed line is for a counterfactual outside the hull. Each line gives the cumulative distribution of our modified Gower distance measures. Take, for example, the value of G^2 (given horizontally) and our rule of thumb of one geometric variability. In this case, this is approximately 0.1, which is an average distance of 10% from the minimum to the maximum values on each variable in X.[12] Essentially no real country-years are within 0.10 or less of this counterfactual for changing Saudi Arabia to a democracy, but about 25% of the data are within this distance for Hungary. Similarly, just a few observations in the data are within even 0.15 of Canada changing to an autocracy, although about a quarter of the country-years are within this distance for Poland. Of course, the full cumulative densities in the figure provide more information than this one point.

We now examine a larger set of counterfactuals all at once with more convenient numerical summaries along the lines of our verbal description of Fig. 3. We start with all variables set at their actual values and then ask what would happen to all autocracies if

[12]The exact geometric variability in the data is 0.1176.

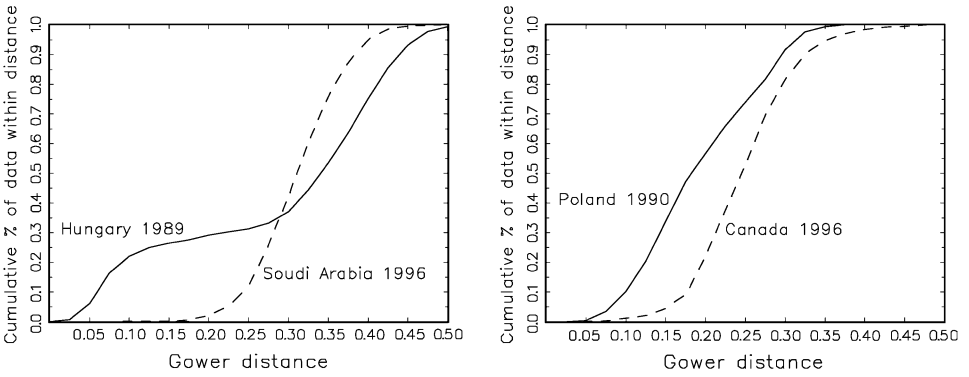

Fig. 3 Distance to four counterfactuals: Cumulative frequencies of modified Gower distances. Countries in the left graph are changed from autocracies to democracies, and the reverse in the right graph. Dashed lines are outside the convex hull of observed data; solid lines are within it.

they became full democracies, and to all full democracies if they became autocracies. This analysis includes 5,814 country-years, with 1,775 full democracies and 4,039 autocracies. What we found was that only 28.4% of the country-years in this widely examined counterfactual fell within the convex hull of the observed data. This means that to analyze this counterfactual in practice, 71.6% of the country-years would require extrapolation and would thus risk a high level of model dependence regardless of the model applied or dependent variable analyzed. As Table 1 summarizes, the result is not symmetric: Among the full democracies switched to autocracies, 53% require interpolation, whereas among the autocracies switched to full democracies, only 17% are interpolation problems. Unfortunately, little discussion in the literature reflects these facts.

The last two columns of the table provide the fraction of countries within a modified Gower distance of 0.1 of a counterfactual, averaged over all counterfactuals for a given type of change in democracy. For example, across the 4,039 country-years where we could hypothetically change autocracies to partial democracies, an average of only 4.2% of the data points are this close to the counterfactual. This is of course considerably better than 0%, but it effectively reduces the real empirical content of the data in making this counterfactual inference to only a small number of observations. The Gower distance test then reveals that 95.8% of the original data are of little help in drawing inferences.

The overall picture in this table is striking. Studying the effects of changes in democracy has been a major project within comparative politics and international relations for at least half a century. This table applies to almost every such analysis with democracy as an explanatory variable in any field with the same or similar control variables, regardless of the choice of dependent variable. Some areas and counterfactuals are less strained than others, but the results here show that most inferences in these fields (or results on most countries within each analysis) are highly model dependent, based much more on unverifiable assumptions about the model than on empirical data. A large fraction are highly model-dependent extrapolations, but even those that are interpolations are fairly distant from available data.

3 Causal Inference

We now turn to counterfactual evaluation as part of causal inference. We start with a definition of causal effects, then our decomposition of the bias in estimation, and finally

Table 1 How factual are counterfactuals about democracy?

			Average % of Data "Nearby"	
Counterfactuals	*N*	*% in Hull*	*All*	*In Hull only*
Full Democracy to Autocracy	1775	53.1	5.5	8.4%
Autocracy to Full Democracy	4039	17.6	2.4	8.2
Partial Democracy to Autocracy	1376	80.5	12.3	14.7
Autocracy to Partial Democracy	4039	61.8	4.2	6.0

a discussion of the components of bias. We devote the most space to discussing the component of bias due to extrapolation, during which we show how the techniques introduced in section 2 can also help solve an existing problem in causal inference.

3.1 *Causal Effects Definition*

To fix ideas, we use a version of the democratic peace hypothesis as a running example, which states that democratic dyads are less conflictual than other dyads (with our discussion generalizing to all possible dependent variables). Let D denote the "treatment" (or "key causal") variable where $D = 1$ denotes a democratic dyad and $D = 0$ denotes a nondemocratic dyad.[13] The dependent variable is Y, the degree of conflict.

To define the causal effect of democracy on conflict, we denote Y_1 as the degree of conflict that would be observed if the dyad were democratic and Y_0 as the degree of conflict otherwise. Obviously, only either Y_0 or Y_1 is observed for any one country at any given time, but not both, since (in our present simplified formulation) a dyad either is or is not democratic. That is, we observe only $Y = Y_0(1 - D) + Y_1D$.

In principle, the democracy variable can have a different causal effect for every dyad in the sample. We can then define the causal effect of democracy by averaging over the whole world, or for the democratic and nondemocratic dyads separately (or for any other subset of dyads). For democratic dyads, this is known as the "average causal effect among the treated," which we define as follows:

$$\begin{aligned}\theta &= \mathrm{E}(Y_1 \mid D = 1) - \mathrm{E}(Y_0 \mid D = 1) \\ &= \text{Factual} - \text{Counterfactual}\end{aligned} \tag{11}$$

We call the first term factual since Y_1 is observable when $D = 1$, although the expected value still may need to be estimated. We refer to the second as counterfactual because Y_0 (the degree of conflict that would exist in a dyad if it were not democratic) is not observed and indeed is unobservable in democratic dyads ($D = 1$). The causal effect for nondemocratic dyads ($D = 0$) is directly analogous and also involves factual and counterfactual terms.

Although medical researchers are almost always interested in θ, political scientists are also interested in the average causal effect for the entire set of observations,

$$\gamma = \mathrm{E}(Y_1) - \mathrm{E}(Y_0), \tag{12}$$

[13]The analysis of treatments with more than two levels follows analogously (e.g., Imai and Dyk 2004; Lechner 1999). We focus on the binary case for expository purposes.

where both terms have a counterfactual element, since each expectation is taken over all dyads, but Y_1 is observed only for democratic dyads and Y_0 only for nondemocratic dyads. These definitions of causal effects are used in a wide variety of literatures (Rubin 1974; Holland 1986; King et al. 1994; Robins 1999a, 1999b; Pearl 2000).

A counterfactual x in this context therefore takes the form of some observed data with only *one* element changed—for example, the Mexico-Spain dyad with all its attributes fixed but with the regime type in both changed to autocracy. Of course, we can easily evaluate how reasonable it is to ask about this counterfactual in one's data with the methods already introduced in section 2: by checking whether x falls in the convex hull of the observed X and computing the distance from x to X. In addition, since x has only one counterfactual element we show that we can easily consult another criterion, whether x falls on the *support* of X, although we discuss some problems with this alternative in section 3.6.[14]

In real applications, the true causal effect, θ or γ, is unknown and needs be estimated, often from observational data since social experiments are costly or, in the case of our data, infeasible. In section 3.2, we discuss the sources of potential problems in using observational data to estimate these causal effects. We focus on θ there for expository purposes as is usual in the statistical literature. However, unlike prior literature, we have generalized our proofs (in Appendix B) to show that our results also hold for the effect on nondemocracies and for the overall average treatment effect, γ, as well. Our empirical examples analyze the overall average causal effect, which is the usual parameter of interest in political science. In addition to illuminating sources of potential problems in causal inference, the estimation bias decomposition shows that inference involving counterfactuals not on the support of X is a critical source of bias, and the present methods of assessing support are often inadequate to the task.

3.2 *Bias Decomposition*

We begin with the simplest estimator of θ using observational data, the difference in means (or, equivalently, the coefficient on D from a regression of Y on a constant and D):

$$\begin{aligned} d &= \text{mean}(Y \mid D = 1) - \text{mean}(Y \mid D = 0) \\ &= \text{mean}(Y_1 \mid D = 1) - \text{mean}(Y_0 \mid D = 0), \end{aligned} \tag{13}$$

where $\text{mean}(a) = \sum_i a_i/n$ (for any vector a with elements a_i for $i = 1, \ldots, n$). The first line is the data-based analogue to Eq. (11), whereas the second recognizes that, for example, when $D = 1$, $Y = Y_1$. To identify the sources of potential problems using observational data in causal inference, we now present a new decomposition of the bias $\text{E}(d - \theta)$ of d as an estimator of the causal effect θ. This decomposition generalizes the three-part decomposition of Heckman, Ichimura, Smith, and Todd (1998). Their decomposition was applied to a simpler problem that does not adequately represent the full range of issues in causal inference in political science. Our new version helps to identify key threats to causal inference in our discipline, as well as to focus on where counterfactual inference is most at issue. In addition to identifying another key component of bias, we present the decomposition for both quantities of interest, γ and θ, whereas Heckman,

[14]The support of X is the range of values of X that are possible (i.e., have positive density) whether or not they occur in our data (see also note 9).

Ichimura, Smith, and Todd (1998) derived the result only for the latter. Both results appear in Appendix B. Thus, for θ, we show that

$$\begin{aligned}\text{bias} &\equiv \mathrm{E}(d-\theta)\\ &= \mathrm{E}[\text{mean}(Y_1 \mid D=1) - \text{mean}(Y_0 \mid D=0) - \theta]\\ &= [\mathrm{E}(Y_1 \mid D=1) - \mathrm{E}(Y_0 \mid D=0)] - [\mathrm{E}(Y_1 \mid D=1) - \mathrm{E}(Y_0 \mid D=1)]\\ &= \mathrm{E}(Y_0 \mid D=1) - \mathrm{E}(Y_0 \mid D=0) \qquad (14)\\ &= \Delta_o + \Delta_p + \Delta_i + \Delta_e. \qquad (15)\end{aligned}$$

We derive the last equality and give the mathematical definition of the terms Δ_o, Δ_p, Δ_e, and Δ_i in Appendix B. These four terms denote exactly the four sources of bias in using observational data, with the subscripts being mnemonics for the components (i.e., the equation is mathematically accurate and not merely an informal analogy). The bias components are due to, respectively, omitted variable bias (Δ_o), post-treatment bias (Δ_p), interpolation bias (Δ_i), and extrapolation bias (Δ_e). Briefly, Δ_o is bias due to omitting relevant variables such as common causes of both the treatment and the outcome variables; Δ_p is bias due to controlling for the consequences of the treatment; Δ_i is bias that can result if not properly adjusting for included controls within the region of the observed data; Δ_e is bias that can occur when extrapolating beyond the range of data in adjusting for included controls. We now explain and interpret each of these components in more detail with particular focus on extrapolation bias, including a discussion of how to use the methods we developed in section 2 to help identify extreme counterfactuals in causal inference.

3.3 *Omitted Variable Bias*

The absence of all bias in estimating θ with d would be assured if we knew (from Eq. [14]) that

$$\mathrm{E}(Y_0 \mid D=1) = \mathrm{E}(Y_0 \mid D=0). \qquad (16)$$

Assumption (16) says that it is safe to use the observed control group outcome ($Y_0 \mid D=0$, the level of conflict initiated by nondemocracies) in place of the unobserved counterfactual ($Y_0 \mid D=1$, the level of conflict initiated by democracies, if they were actually nondemocracies.)

Since valid inference without controls is rarely the case, we introduce control variables: Let Z denote a vector of control variables (explanatory variables aside from D), and denote the set of all explanatory variables as $X = \{D, Z\}$. If, after controlling for Z, treatment assignment of D is random—that is, if we measure and control for the right set of control variables (such as those that are common causes of D and Y), so that

$$\mathrm{E}(Y_0 \mid D=1, Z) = \mathrm{E}(Y_0 \mid D=0, Z) \qquad (17)$$

holds, then from Eq. (27) in Appendix B, the first component of bias vanishes: $\Delta_o = 0$. Thus this first component of bias, Δ_o, is due to pertinent control variables being omitted from X so that Eq. (17) is violated. This is the familiar omitted variable bias, which can plague any model. It can also be due to controlling for irrelevant variables in certain situations, so Z should be minimally sufficient (Greenland et al. 1999).

Since endogeneity bias and selection bias can be written as omitted variable bias, Δ_o encompasses these problems as well. To be specific, endogeneity bias, selection bias, and omitted variable bias each cause inferential problems by inducing a correlation between the explanatory variables and the error term. If we control for the correct variables, then it is sometimes possible to eliminate these problems. With omitted variable bias, the controls to include would be omitted variables that are common causes of D and Y. Similarly, we can avoid the biases due to nonrandom selection if we control for the probability that each unit is selected into the sample, and we can eliminate endogeneity bias by including in the controls covariates that eliminate the conditional relationship between X and the error term.

3.4 *Post-treatment Bias*

The second component of bias in our decomposition, Δ_p, deviates from zero when some of the control variables Z are at least in part consequences of the key causal variable D. If Z includes these post-treatment variables, then when the key causal variable D changes, the post-treatment variables may change too. Thus, if we denote Z_1 and Z_0 as the values that Z takes on when $D = 1$ and $D = 0$, respectively (components of Z that are strictly pre-treatment do not change between Z_0 and Z_1), then, as with Y, either Z_1 or Z_0, but not both, are observed for any one observation, and the observed $Z = Z_0(1 - D) + Z_1(D)$ will be different from the counterfactual Z_0, resulting in a nonzero Δ_p in Eq. (24).

As a simple example that illustrates the bias of controlling for post-treatment variables, suppose we are predicting the vote with partisan identification. If we control for the intended vote five minutes before walking into the voting booth, our estimate of the effect of partisan identification would be nearly zero. The reason is that we are inappropriately controlling for the consequences of our key causal variable, and for most of the effects of it, thus biasing the overall effect. Yet we certainly should control for a pre-treatment variable like race that cannot be a consequence of partisan identification but may be a confounding variable. Thus causal models require separating out the pre- and post-treatment variables and controlling only for the pre-treatment, background characteristics.

To avoid this component of bias, Δ_p, we need to ensure that we control for no post-treatment variables, or at least that the distribution of our post-treatment variables does not vary with D:

$$\mathrm{P}(Z \mid D = 1) = \mathrm{P}(Z \mid D = 0), \tag{18}$$

so that $Z_0 = Z_1 = Z$. If this assumption holds, then $\Delta_p = 0$ in Eq. (24) vanishes.

Post-treatment variable bias is a large and often overlooked component of bias in estimating causal effects in political science (see King 1991; King et al. 1994, pp. 173ff). It is known in the statistical literature but is assumed away in most models and decompositions (Frangakis and Rubin 2002). This decision may be reasonable in other fields, where the distinction between pre- and post-treatment variables is easier to recognize and avoid, but in political science, especially comparative politics and international relations, the problem is often severe. For example, is GDP a consequence or cause of democracy? How about educational levels? Fertility rates? Infant mortality? Trade levels? Are international institutions causes or consequences of international cooperation? Many or possibly even most variables in these literatures are both causes and consequences of whatever is regarded as the treatment (or key causal) variable. Thus the fundamental problem with much research in comparative politics and international relations is not merely the bias induced by controlling for post-treatment variables. The problem is that even if dropping out these variables alleviates post-treatment bias, it will likely also induce omitted variable bias.

In our field, unfortunately, we almost always need to consider both Δ_o and Δ_p together, and in many situations we cannot fix one without making the other worse. The same is not true in all fields (which is perhaps the reason the Δ_p component was ignored by Heckman, Ichimura, and Todd 1998), but it is rampant in ours. Unfortunately, the news gets worse, since even the methodologist's last resort—try it both ways and, if it doesn't make a difference, ignore the problem—does not work here. Rosenbaum (1984, pp. 664ff) studies the situation in which we run two analyses, one including and one excluding the variables that are partly consequences and partly causes of X. He shows that the true effect could be greater than these two or less than both. It is hard to emphasize sufficiently the seriousness of this problem and how prevalent it is in comparative politics and international relations.

Although we have no general solution to this problem, we can offer one way to avoid both Δ_p and Δ_o in the presence of variables that are partially post-treatment. Aside from choosing better research designs in the first place, of course, our suggestion is to study *multiple-variable causal effects*. If we cannot study the effects of democracy controlling for GDP because higher GDP is in part a consequence of democracy, we may be able to study the joint causal effect of a change from nondemocracy to democracy *and* a simultaneous increase in GDP. This counterfactual is more realistic, i.e., closer to the data, because it reflects changes that actually occur in the world and does not require us to imagine holding variables constant that do not stay constant in nature. If we have specified a parametric model with both variables, we can study this question by simultaneously moving both GDP and democracy while holding constant other variables. An alternative would be to recode the two variables into one on, as much as possible, a single dimension.

If this alternative formulation provides an interesting research question, then it can be studied without bias due to Δ_p since the joint causal effect will not be affected by post-treatment bias. Moreover, the multiple-variable causal effect might also have no omitted variable bias Δ_o, since both variables would be part of the treatment and could not be potential confounders. Of course, if this question is not of interest and we need to stick with the original question, then no easy solution exists at present. At that point, we should recognize that the counterfactual question being posed is too unrealistic and too strained to provide a reasonable answer using the given data with any statistical model. Either way, this is a serious problem that needs to move higher on the agenda of political methodology.

3.5 *Interpolation Bias*

For clarity, we now assume that the two components of bias we have previously discussed are not an issue, so we have no post-treatment bias, and we have the minimally sufficient set of pre-treatment variables Z to control for. However, even when $\Delta_o = 0$ and $\Delta_p = 0$, we still have to control for Z properly. The two remaining components of bias—interpolation bias and extrapolation bias—both have to do with correctly identifying the necessary control variables but failing to adjust for them properly. Interpolation bias or Δ_i results from incorrect adjustment for control variables in regions of interpolation, and extrapolation bias results from improperly adjusting for controls where data are needed but do not exist.

Interpolation bias may exist in the simple difference in means estimator if the measured control variables Z are related in any way to the treatment variable, that is, if the multivariate density of Z for the treatment group differs from that for the control group (within the region of interpolation). If in addition to these density differences Z also affects the outcome variable, then interpolation bias will exist if the density differences in Z are not properly adjusted.

When using a parametric model to adjust for control variables, this component of bias arises from controlling for Z with the wrong functional form. For example, in an application without post-treatment bias, with all control variables that could cause bias identified, and where extrapolation is unnecessary, our estimator could still generate bias by choosing a linear model to adjust for controls if the data were generated from a quadratic. Fortunately, standard regression diagnostics are quite useful for checking model fit within the range of the data.

Interpolation bias can also be adjusted for without a specified functional form via matching, inverse propensity score weighting, or nonparametric smoothing (e.g., Rosenbaum and Rubin 1984; Heckman, Ichimura, and Todd 1998; Robins 1999a, 1999b; Winship and Morgan 1999). Ultimately, whatever method of adjustment is used, the two multivariate densities of Z for the control and treatment groups need to be the same for interpolation bias to be eliminated. We provide further insight into interpolation bias during our discussion of extrapolation bias, to which we now turn.

3.6 *Extrapolation Bias*

The last component of bias, and the one most related to the central theme of this article, is extrapolation bias. This component is the second of the two that arise from not adjusting or improperly adjusting for identified control variables, but since it occurs beyond the range of the observed data, avoiding extrapolation bias is potentially much more difficult (and, as we showed above, model dependent) than avoiding interpolation bias.

From Eq. (25), we see that extrapolation bias may arise when the support of the distribution of Z for the treatment group differs from that of the control group. That is, there may be certain values of Z that some members of one group take on with positive probability but no members of the other group possess. For example, we might observe no full democracies with GDP as low as in some of the autocracies but still somehow need to control for GDP. Intuitively, these autocracies have no comparables in the data and so are not readily useful for estimating causal effects. To make causal inferences in situations with nonoverlapping densities, we must therefore either eliminate the region outside of common support (as is standard practice in statistics and medicine) or attempt to extrapolate to the needed data (e.g., autocracies with high GDP), such as by using a parametric model (as is standard practice in political science and most of the other social sciences). As we demonstrated in section 2, extrapolation in forecasting involves considerable model dependence. The same issue applies in causal inference, as we discuss below. Thus, unless we happen to be in the extraordinary situation in which a known theory or prior evidence makes it possible to narrow down the possible models to one, or we happen to guess the right model, we will be left with extrapolation bias, $\Delta_e \neq 0$.

We begin with a simple illustration of extrapolation bias in hypothetical data with a single control variable. We then discuss the idea of and problems with the propensity score approach commonly used to identify regions of extrapolation in applications with more than one control variable. Finally, we show how our convex hull approach can be used to help solve problems with the propensity score approach in many cases and to assist in others.

3.6.1 Illustration with a Single Control Variable

Figure 4 illustrates some key issues involved in data that generate the need to extrapolation in causal inference. The figure also illustrates the connection between the problems of

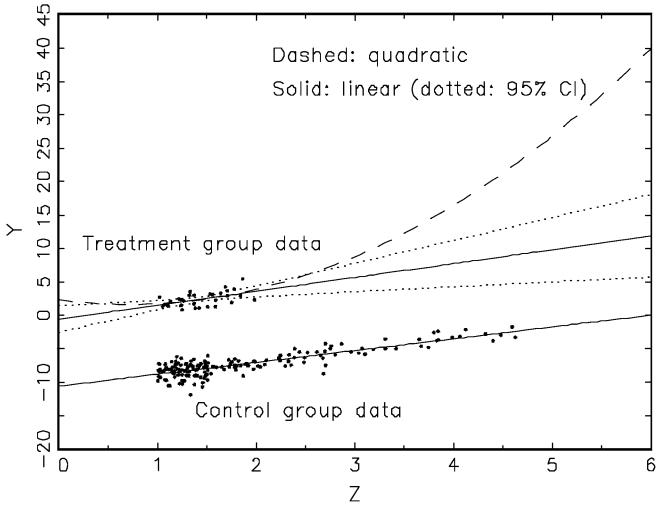

Fig. 4 An illustration of how the degree of extrapolation bias is more severe (and model dependent) than interpolation bias.

extrapolation in causal inference and extrapolation in forecasting and what-if questions discussed earlier. Figure 4 plots hypothetical data on the dependent variable vertically and a single control variable Z horizontally. The treatment and control groups are labeled and the points are clearly separated in the figure. To estimate the causal effect in these data, we make comparisons between the treatment and control groups on the vertical axis (which corresponds to the outcome variable). The key extrapolation problem is that there exist no treated units for values of $Z > 2$ where some control data do exist, and so any comparison between the treated and control groups in this region would be based on extrapolating the treatment group data from where it is observed to where it is needed. In other words, a study seeking to estimate a causal inference from data where extrapolation is necessary has the same problem in that region as not having a treatment (or control) group at all.

As the figure shows, the two models fitted to the treated data, one linear and one quadratic, fit the treated data almost identically, but in the region to which the counterfactual extrapolations are needed (i.e., where control units exist but treated units do not), the difference between the models is vast. This illustrates model dependence, of course, but it also illustrates extrapolation bias, since at least one of the models shown must be false in the extrapolation region, so if used it would generate bias and make $\Delta_e \neq 0$. Since we have no data to test which model is appropriate, or whether both are wrong in the extrapolation region, we have no means to rule out extrapolation bias based on empirical evidence.

Interpolation bias could be seen in the figure if the different functional forms fitted to the treated data differed in the sample. If that were the case (and it is not as drawn), then bias would result if the estimation model were not close to the model that represented the data. In practice, because model dependence is much less of an issue in areas of interpolation (or on the common support) than in areas of extrapolation, interpolation bias can often be detected and corrected in ways that extrapolation bias cannot.

If we use the data outside the region of common support, we must extrapolate and will therefore have some degree of model dependence and thus risk some bias for almost any model chosen. Alternatively, we can delete nonoverlap data, which eliminates the need to

extrapolate. Of course, this procedure would fail to produce any estimates at all in applications where no data lie on the common support, a problem with some prevalence in our field. If some data do lie within the common support region and the quantity of interest is the average treatment effect (γ in Eq. [12]), dropping observations outside of common support will produce bias by definition, as it changes the population and thus the quantity of interest. Similarly, in the situation in which we convince ourselves that we are interested only in the average treatment effect on the treated (θ in Eq. [11]), dropping treatment units not on common support will result in bias by changing the population of inference.[15]

Although extrapolation bias is hard to correct without access to better data or willingness to change the population of inference (and thus the research questions), identifying the regions of extrapolation is important in all applications. It may be disappointing, of course, to know that the desired questions have no good answers in available data, but it is better to know this than to ignore it.

3.6.2 Identification of Extrapolation Regions

We now turn to the task of identifying regions of extrapolation in causal inference, which are regions outside of the common support of Z for the treatment and control groups. We noted earlier that being on the support of X is another criterion for checking the quality of a counterfactual. It is easy to see that for causal inference, the counterfactual x implied in the second term of Eq. (11) being on the common support of $Z \mid D = 1$ and $Z \mid D = 0$ is in fact equivalent to being on the support of $X \equiv \{D, Z\}$. Thus identifying the regions of common support is also a direct check of the quality of the key counterfactual required for causal inference (directly analogous to section 2.3). In the simple case in which Z contains just one variable, we can simply plot both histograms on the same horizontal scale and compare them. Areas requiring extrapolation can easily be identified from the histograms as the areas that do not overlap. Interpolation bias can arise where the histograms overlap but differ in density.

In most real applications, of course, Z contains many control variables, and so comparing features of $P(Z \mid D = 1)$ and $P(Z \mid D = 0)$ would involve estimating and comparing two multidimensional densities. For more than a few explanatory variables, this is a formidable task (essentially impossible without stringent assumptions). Below we briefly review one popular approach to this curse of dimensionality problem, illuminate some serious difficulties of that approach for our task, and then discuss the utility of our methods introduced in section 2 as an alternative approach.

Difficulties with the Propensity Score Approach. One approach that has been used to compare multivariate Z distributions and assess multivariate common support in causal inference problems uses the *propensity score*, $\pi \equiv \Pr(D = 1 \mid Z)$, the probability of $D = 1$ given the control variables Z. The propensity score summarizes the multidimensional Z with a unidimensional π. Rosenbaum and Rubin (1983) show that conditioning on the *true* π balances the distribution of Z across the treatment and control groups:

$$\mathrm{P}(Z \mid D = 1, \pi) = \mathrm{P}(Z \mid D = 0, \pi). \tag{19}$$

[15] If θ is the quantity of interest and if all treatment units are on common support, dropping control group data not on the common support will lead to inefficient estimates.

This statement is of great theoretical interest, since it makes it possible to demonstrate that the condition for ensuring elimination of both control and extrapolation bias, that the distribution of Z in the treatment group being the same as that in the control group:

$$\mathrm{P}(Z \mid D = 1) = \mathrm{P}(Z \mid D = 0), \tag{20}$$

is equivalent to a seemingly more useful form, that the distribution of π in the treatment group is identical to the distribution of π in the control group:[16]

$$\mathrm{P}(\pi \mid D = 1) = \mathrm{P}(\pi \mid D = 0). \tag{21}$$

This expression seems more useful because it apparently solves the curse of dimensionality problem in the multivariate comparison in Eq. (20) by only requiring comparison of two unidimensional densities of π (one for democracies and one for nondemocracies) and identifying common support and adjusting the density differences based only on this one control variable.

The key problem, however, is that the "true" propensity score is itself unknown and must be estimated from the data, so any estimate of π may be afflicted not merely by sampling error but also potentially by the usual model specification, measurement, and other estimation errors. If the estimated propensity score is wrong, then the theoretical properties of the propensity score do not hold, and no result suggests that it will still give valid results. Thus the curse of dimensionality is not solved by the propensity score approach in real applications. And since the only way to check whether an estimated propensity score is close to being true is to check whether it balances the Z distributions, the theory of the propensity score from the point of view of the researcher is tautological (Ho et al. 2005).

For the task of identifying the extrapolation region or nonoverlap in the support of $Z \mid D = 1$ and $Z \mid D = 0$, using the propensity score involves not only what (Ho et al. 2005) call the Propensity Score Tautology, but also a fundamental problem of infinite regress: we cannot use the propensity score to identify regions of extrapolation until we can verify that the estimated propensity score is valid, but we cannot verify the validity of the estimated propensity score until we have first removed the regions requiring extrapolation. The reason is that by definition, π values outside of common support will not satisfy the balancing condition (19), which is the basis of all balancing tests, since one of the two densities will be 0 and the other positive. For estimated propensity scores, the question is whether the imbalance is due to incomplete overlap of the $Z \mid D = 1$ and $Z \mid D = 0$ distributions or to the inadequacy of the model used for estimating the propensity score. Since it is virtually impossible to prove that any given propensity score model is close to being the true model in the presence of imbalance, the researcher will not be able to differentiate the two different causes of the imbalance. Thus, in theory, to identify nonoverlap with an estimated propensity score would require the *assumption* that the propensity score model is correct.[17] Of course, verification by assumption is obviously not an empirical exercise, so using the propensity score for identifying the extrapolation region is of questionable practical value.

[16] By definition, $\mathrm{P}(Z \mid D = 0) = \int \mathrm{P}(Z \mid D = 0, \pi) P(\pi \mid D = 0) d\pi$, which upon substituting in Eq. (19), gives $\mathrm{P}(Z \mid D = 0) = \int \mathrm{P}(Z \mid D = 1, \pi) \mathrm{P}(\pi \mid D = 0) d\pi$. Comparing this result with $\mathrm{P}(Z \mid D = 1) = \int \mathrm{P}(Z \mid D = 1, \pi) \mathrm{P}(\pi \mid D = 1) d\pi$, which is also true by definition, proves that Eq. (20) holds if Eq. (21) does.

[17] Standard balancing tests are not reliable because they reduce the comparison of two densities at a given π to comparison of a few moments of the densities within intervals of π values. In hard balancing problems or with sparse data, wider intervals are typically used or the comparison is restricted to the overlap region, further undermining the value of the tests.

The attractive theoretical properties of the true propensity score have sparked much interest, and the propensity score methodology has been used in numerous applications throughout a wide array of disciplines. But as our discussion makes clear, using the propensity score to identify common support or the extrapolation region is difficult at best, and likely misleading in many situations. Even if a researcher has good reason to use the propensity score to help in matching, weighting, or parametric analyses, having a method to identify, and possibly remove observations from, the region of extrapolation first would help not only in removing extrapolation bias but in making balancing checks more reliable in evaluating estimated propensity scores.

Using the Convex Hull. Fortunately, based on our analyses in the first part of this article, a workable approach to this problem is available. If we are interested in estimating the average treatment effect on the treated (θ in Eq. [11]), then we simply discard any control units for which Z is not within the convex hull of the treated units Z. (Even if some of the treated units are outside the convex hull of the control units and thus would require extrapolation, they would not be omitted so the quantity of interest remains the same, although it would be worth identifying them so a source of the remaining model dependence is identified.)

If instead we are wiling to change the quantity being estimated to something different but reliably estimate without high levels of model dependence, we would also want to drop treated units that fall outside the convex hull of the control units. If this alternative is desired, we can consolidate the two steps and proceed as follows. Let D^* and Z^* denote, respectively, the subset of D and Z such that the counterfactual points $\{1 - D^*, Z^*\}$ fall within the convex hull of the observed data $X \equiv \{D, Z\}$; then use the convex hull of Z^* as an estimate of where the common support lies. When no extrapolation is needed, $Z^* = Z$. *Thus the same procedures for identifying whether points fall within the convex hull as described in section 2.3 can be used for assessing common support*. Both procedures are conservative evaluations of common support and more so in higher dimensional space, but each is fast, easy to apply, and applicable to a wide range of problems.

In essence, this strategy uses the convex hull of the data as an estimate of the support of the data. In sufficiently large samples, the convex hull contains the support; when the support has no gaps or voids, the convex hull approximation is nearly exact. To avoid the risk of voids within the common support, we can use the Gower distance to assess whether any of the counterfactual points within the hull are far from any observed data.[18]

In small samples, the convex hull may give too narrow a range for the common support region (for the same reason that the sample maxima and minima are biased estimates of the corresponding population quantities, for any finite sized sample). However, where only one sample from the same population will ever be observed, as in most areas of comparative politics or international relations, no other observable data will be available to differentiate the convex hull boundary from that of the support, and using the convex hull will still normally be a reasonable choice.

This strategy has not been used in the literature before, in part because finding whether counterfactual points fall in the hull has not previously been viewed as feasible. Given our methods described in section 2.3.2, however, this strategy is now feasible and easy to apply. Indeed, a key advantage of the strategy suggested here is that what is at least a good first cut at finding the region of common support can now be automated and easily included

[18]If the risk of voids seems substantial enough, it would be advisable as an alternative or extra verification step to rule counterfactual points not within, say, 0.1 Gower distance from a sufficient number of data points as effectively off the common support, although in many applications checking the convex hull should be sufficient on its own.

in standard statistical software. It is already included in the software that accompanies this (Stoll et al. 2005) and has also been implemented as part of a general purpose matching software package called MatchIt (Ho et al. 2005).

4 Concluding Remarks

Consider a model that fits the data well, has large coefficients, small p-values, narrow confidence intervals, large causal effect estimates, predictions with path-breaking policy implications, and fascinating answers to a range of what-if questions. With statistical reporting standards now commonly used in political science, results like these would be written up, published, and taken seriously by readers. Unfortunately, a subset of these involve counterfactuals that are so model dependent as to be nearly unrelated to the data at hand and so are based more on the authors' hypotheses and convenient but undefended or unnoticed model assumptions than the data. Although it is rarely the case presently, assessing model dependence of counterfactual questions needs to be a routine and expected part of statistical reporting for anyone making predictions, asking what-if questions, and estimating causal effects—which together encompasses the goals of a large fraction of empirical work in the discipline.

We have offered several approaches to evaluating whether counterfactuals are too far from the data that are easy to use and should generally be consulted prior to drawing substantive conclusions. Other approaches such as sensitivity analysis, Bayesian model averaging, committee methods, or transdimensional Markov Chains are also useful, but only when it is feasible to identify the relevant class of models for exploration (Hoeting et al. 1999; Imai and King 2004; Sisson 2005) In most situations, the approach we recommend, which does not require choosing classes of models or specifying or estimating any models at all, should be of wider applicability and have greater power.

Our checks on model dependency hold regardless of the model chosen and for all possible dependent variables. However, they are conditional on the choice of explanatory variables and their valid measurement. Measurement error and the incorrect identification of relevant covariates must be avoided in these procedures, as in all others. Passing convex hull and Gower distance tests can help in assessing the degree of model dependence, but all the other usual threats to validity must still be evaluated and avoided.

If an interesting counterfactual question is so far from the data that answers are highly model dependent, we still may wish to draw conditional inferences that are by their nature more uncertain than model-based uncertainty measures indicate. The best one could do in that situation would be to fairly indicate the degree of model dependence when reporting results. If changing the counterfactual question at hand is not an option, then the most substantively productive procedure would be to design research to avoid the problem from the start. When this turns out to be possible, it basically involves collecting data more relevant to the question at hand. Accomplishing this is often much easier at the research design stage than during data analyses.

Appendix A: Membership in a Convex Hull as a Linear Programming Problem

This appendix gives our approach to checking whether a given point $x_{1\times k}$ is in the convex hull of $X_{n\times k}$, where n is the number of data points in X and k the number of variables. Let S be the set of vertices of the convex hull of X, containing all the "boundary" points of X. By definition, x being in the convex hull of X implies that x can be expressed as a convex combination of points in S. Since all points of X are also convex combinations of points in S, the condition is equivalent to x being a convex combination of all points in X. Identification

of S can be computationally very expensive, but the second form of the condition can be checked easily using standard linear programming software without ever computing S.

To do so, we formulate the problem as one of checking the existence of a feasible solution for a standard linear programming problem with a degenerate objective function. If x can be expressed as a convex combination of points in X, then there exists a vector of coefficients $\eta_{n\times 1}$ constrained to the simplex so that $X'\eta = x'$. This last equation contains k linear constraints, each stating that an element (variable) of x is a convex combination of the corresponding elements of rows in X. Combining this with the constraint that the elements of η sum to one, we have a total of $k + 1$ linear constraints in the form $A'\eta = B'$, where A' and B' are X' and x' with a row of ones added, respectively.

To check whether x is in the convex hull of X therefore is equivalent to checking the existence of a feasible solution to the following standard form linear programming (LP) problem:

$$\begin{aligned} &\min C'\eta \\ &\text{s.t. } A'\eta = B' \\ &\qquad \eta \geq 0, \end{aligned} \tag{22}$$

where C is a vector of zeros (so there is no objective function to minimize). Checking whether there is a feasible solution to problem (22) is what all standard LP software does in phase I, and it can be done very efficiently for large k and n.

Appendix B: Decomposition of Causal Effect Estimation Bias

We now derive our new decomposition of the bias of the estimator d in Eq. (13). We derive this for the average treatment effect and for the average treatment effect on the treated. Note that d is the simplest estimator for all three causal effect parameters (the average causal effect in democracies, θ, its counterpart for nondemocracies for which we have assigned no symbol, and the average causal effect overall, γ, from Eq. [12]). In this appendix we prove that the bias of d as an estimator of any of these three parameters has the same four types of components as given in Eq. (15) and discussed in section 3.2.

The first two parts of our decomposition, Δ_o and Δ_p, correspond to, or in a sense provide another proof for, Pearl's "back-door" criterion for identifying adjustment variables in the identification of causal effects. This criterion has two parts: that Z should not be caused by X and that Z should be minimally sufficient to control for omitted variable bias. The first two components of our decomposition are analogous, but in reverse: we show in estimation that the bias would be zero if X did not cause Z, and Z includes the right variables. Since he was concerned with identification and not estimation, Pearl did not explicitly analyze the consequences of nonoverlap or density differences, although he implicitly assumed their absence. In contrast, Heckman, Ichimura, and Todd's (1998) three-part decomposition, which we generalize here, does not address or analyze the consequences of post-treatment bias, except implicitly, which is fundamental in Pearl's work and essential to understanding a key problem in the comparative politics literature.

We start by showing that the bias of d in estimating the total effect γ is a convex combination of its bias in estimating the two group-specific causal effect parameters. We have $\mathrm{E(d-\gamma)=[E(Y_1 \mid D=1)-E(Y_0 \mid D=0)]-[E(Y_1)-E(Y_0)]}$. Let $\tau = \Pr(D = 1)$ be the size of the treatment group and then rewrite the terms in the definition of γ above as $\mathrm{E}(Y_1) = \tau\mathrm{E}(Y_1 \mid D = 1) + (1 - \tau)\mathrm{E}(Y_1 \mid D = 0)$ and $\mathrm{E}(Y_0) = \tau\mathrm{E}(Y_0 \mid D = 1) + (1 - \tau)\mathrm{E}(Y_0 \mid D = 0)$. Thus,

$$\begin{aligned}
\mathrm{E}(d-\gamma) &= \mathrm{E}(Y_1 \mid D=1) - \mathrm{E}(Y_0 \mid D=0) - \tau\mathrm{E}(Y_1 \mid D=1) \\
&\quad - (1-\tau)\mathrm{E}(Y_1 \mid D=0) + \tau\mathrm{E}(Y_0 \mid D=1) + (1-\tau)\mathrm{E}(Y_0 \mid D=0) \\
&= (1-\tau)[\mathrm{E}(Y_1 \mid D=1) - \mathrm{E}(Y_1 \mid D=0)] \\
&\quad + \tau[\mathrm{E}(Y_0 \mid D=1) - \mathrm{E}(Y_0 \mid D=0)] \\
&= (1-\tau)B_0 + \tau B_1, \qquad (23)
\end{aligned}$$

where $B_1 = \mathrm{E}(Y_0 \mid D=1) - \mathrm{E}(Y_0 \mid D=0) = \mathrm{E}(d-\theta)$ is the bias of using d to estimate θ, the causal effect on the treated (democracies), as derived in Eq. (14). In a directly analogous way, $B_0 = \mathrm{E}(Y_1 \mid D=1) - \mathrm{E}(Y_1 \mid D=0)$ is the bias of d as an estimator of the causal effect in the control group (nondemocracies). (Note that, quite intuitively, B_1 is a function of unobservables among the treated, and B_0 is a function of unobservables among the untreated.)

We now derive the four components of bias for B_1, and then in an identical fashion for B_0, before we combine them as per Eq. (23). We have:

$$\begin{aligned}
B_1 &= \mathrm{E}(Y_0 \mid D=1) - \mathrm{E}(Y_0 \mid D=0) \\
&= \mathrm{E}_{z_0}[\mathrm{E}(Y_0 \mid D=1, Z_0) - \mathrm{E}(Y_0 \mid D=0, Z_0)] \\
&\quad - \mathrm{E}_z[\mathrm{E}(Y_0 \mid D=1, Z) - \mathrm{E}(Y_0 \mid D=0, Z)] \\
&\quad + \mathrm{E}_z[\mathrm{E}(Y_0 \mid D=1, Z) - \mathrm{E}(Y_0 \mid D=0, Z)] \\
&= [\mathrm{E}_{z_0}\mathrm{E}(Y_0 \mid D=1, Z_0) - \mathrm{E}_z\mathrm{E}(Y_0 \mid D=1, Z)] \\
&\quad + \mathrm{E}_z[\mathrm{E}(Y_0 \mid D=1, Z) - \mathrm{E}(Y_0 \mid D=0, Z)] \\
&= \Delta_p + \mathrm{E}_z[\mathrm{E}(Y_0 \mid D=1, Z) - \mathrm{E}(Y_0 \mid D=0, Z)]
\end{aligned}$$

where

$$\Delta_p = \mathrm{E}_{z_0}\mathrm{E}(Y_0 \mid D=1, Z_0) - \mathrm{E}_z\mathrm{E}(Y_0 \mid D=1, Z) \qquad (24)$$

is the bias due to controlling for post-treatment variables, and $\mathrm{E}_z[\mathrm{E}(Y_0 \mid D=1, Z) - \mathrm{E}(Y_0 \mid D=0, Z)]$ can be further decomposed, following and generalizing the approach in Heckman, Ichimura, Smith, and Todd (1998). Let S_j denote the support of Z in $F(Z \mid D=j)$ for $j=0, 1$ and S the common support. Then

$$\begin{aligned}
&\mathrm{E}_z[\mathrm{E}(Y_0 \mid D=1, Z) - \mathrm{E}(Y_0 \mid D=0, Z)] \\
&= \int_{S_1} \mathrm{E}(Y_0 \mid D=1, Z)dF(Z \mid D=1) - \int_{S_0} \mathrm{E}(Y_0 \mid D=0, Z)dF(Z \mid D=0) \\
&= \left\{ \int_{S_1 \setminus S} \mathrm{E}(Y_0 \mid D=1, Z)dF(Z \mid D=1) + \int_S \mathrm{E}(Y_0 \mid D=1, Z)dF(Z \mid D=1) \right\} \\
&\quad - \left\{ \int_{S_0 \setminus S} \mathrm{E}(Y_0 \mid D=0, Z)dF(Z \mid D=0) + \int_S \mathrm{E}(Y_0 \mid D=0, Z)dF(Z \mid D=0) \right\} \\
&\quad + \left\{ \int_S \mathrm{E}(Y_0 \mid D=0, Z)dF(Z \mid D=1) - \int_S \mathrm{E}(Y_0 \mid D=0, Z)dF(Z \mid D=1) \right\} \\
&= \Delta_e + \Delta_i + \Delta_o
\end{aligned}$$

where, through regrouping the terms,

$$\Delta_e = \int_{S_1 \setminus S} \mathrm{E}(Y_0 \mid D = 1, Z) dF(Z \mid D = 1) - \int_{S_0 \setminus S} \mathrm{E}(Y_0 \mid D = 0, Z) dF(Z \mid D = 0) \tag{25}$$

$$\Delta_i = \int_S \mathrm{E}(Y_0 \mid D = 0, Z) \{dF(Z \mid D = 1) - dF(Z \mid D = 0)\} \tag{26}$$

$$\Delta_o = \int_S \{\mathrm{E}(Y_0 \mid D = 1, Z) - \mathrm{E}(Y_0 \mid D = 0, Z)\} dF(Z \mid D = 1). \tag{27}$$

Combining results proves Eq. (15) and gives:

$$B_1 = \Delta_p + \Delta_e + \Delta_i + \Delta_o. \tag{28}$$

Decomposition of B_0 proceeds identically and we omit the intermediate steps. The results are:

$$B_0 = \Delta_p^0 + \Delta_e^0 + \Delta_i^0 + \Delta_o^0 \tag{29}$$

where

$$\Delta_p^0 = \mathrm{E}_z \mathrm{E}(Y_1 \mid D = 0, Z) - \mathrm{E}_{z_1} \mathrm{E}(Y_1 \mid D = 0, Z_1) \tag{30}$$

$$\Delta_e^0 = \int_{S_1 \setminus S} \mathrm{E}(Y_1 \mid D = 1, Z) dF(Z \mid D = 1) - \int_{S_0 \setminus S} \mathrm{E}(Y_1 \mid D = 0, Z) dF(Z \mid D = 0) \tag{31}$$

$$\Delta_i^0 = \int_S \mathrm{E}(Y_1 \mid D = 0, Z) \{dF(Z \mid D = 1) - dF(Z \mid D = 0)\} \tag{32}$$

$$\Delta_o^0 = \int_S \{\mathrm{E}(Y_1 \mid D = 1, Z) - \mathrm{E}(Y_1 \mid D = 0, Z)\} dF(Z \mid D = 1). \tag{33}$$

Now, to arrive at the decomposition of bias in estimating the total effect γ, we only need to combine Eqs. (28) and (29) as per Eq. (23). Omitting tedious but straightforward intermediate steps, the results are:

$$\mathrm{E}(d - \gamma) = \Delta_p^t + \Delta_e^t + \Delta_i^t + \Delta_o^t \tag{34}$$

where

$$\begin{aligned}
\Delta_p^t &= (1-\tau)\{\mathrm{E}_z\mathrm{E}(Y_1 \mid D=0,Z) - \mathrm{E}_{z_1}\mathrm{E}(Y_1 \mid D=0,Z_1)\} \\
&\quad + \tau\{\mathrm{E}_{z_0}\mathrm{E}(Y_0 \mid D=1,Z_0) - \mathrm{E}_z\mathrm{E}(Y_0 \mid D=1,Z)\} \\
\Delta_e^t &= \int_{S_1\setminus S}\{(1-\tau)\mathrm{E}(Y_1 \mid D=1,Z) + \tau\mathrm{E}(Y_0 \mid D=1,Z)\}dF(Z \mid D=1) \\
&\quad - \int_{S_0\setminus S}\{(1-\tau)\mathrm{E}(Y_1 \mid D=0,Z) + \tau\mathrm{E}(Y_0 \mid D=0,Z)\}dF(Z \mid D=0) \\
\Delta_i^t &= \int_S \{(1-\tau)\mathrm{E}(Y_1 \mid D=0,Z) + \tau\mathrm{E}(Y_0 \mid D=0,Z)\} \\
&\quad \times \{dF(Z \mid D=1) - dF(Z \mid D=0)\} \\
\Delta_o^t &= \int_S (1-\tau)[\mathrm{E}(Y_1 \mid D=1,Z) - \mathrm{E}(Y_1 \mid D=0,Z)] \\
&\quad + \tau[\mathrm{E}(Y_0 \mid D=1,Z) - \mathrm{E}(Y_0 \mid D=0,Z)]dF(Z \mid D=1).
\end{aligned}$$

The four components of bias have the same qualitative interpretations in Eqs. (28), (29), and (34).

References

Bishop, Christopher M. 1995. *Neural Networks for Pattern Recognition*. Oxford: Oxford University Press.

Cuadras, C. M., and J. Fortiana. 1995. "A Continuous Metric Scaling Solution for a Random Variable." *Journal of Multivariate Analysis* 52:1–14.

Cuadras, C. M., J. Fortiana, and F. Oliva. 1997. "The Proximity of an Individual to a Population with Applications to Discriminant Analysis." *Journal of Classification* 14:117–136.

de Berg, Mark, Marc van Krevald, Mark Overmars, and Otfried Schwarzkopf. 1998. *Computational Geometry: Algorithms and Applications*, 2nd rev. ed. New York: Springer.

Esty, Daniel C., Jack Goldstone, Ted Robert Gurr, Barbara Harff, Pamela T. Surko, Alan N. Unger, and Robert S. Chen. 1998. *The State Failure Task Force Report: Phase II Findings*. McLean, VA: Science Applications International Corporation.

Frangakis, Constantine E., and Donald Rubin. 2002. "Principal Stratification in Causal Inference." *Biometrics* 58:21–29.

Gelman, Andrew, and Gary King. 1994. "Party Competition and Media Messages in U.S. Presidential Election Campaigns." In *The Parties Respond: Changes in the American Party System*, ed. L. Sandy Maisel. Boulder, CO: Westview, pp. 255–295. (Available from http://gking.harvard.edu/files/abs/partycomp-abs.shtml.)

Gower, J. C. 1966. "Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis." *Biometrika* 53(3/4):325–388.

Gower, J. C. 1971. "A General Coefficient of Similarity and Some of Its Properties." *Biometrics* 27:857–872.

Greenland, Sander, Judea Pearl, and James M. Robins. 1999. "Causal Diagrams for Epidemiologic Research." *Epidemiology* 10(1):37–48.

Hastie, Trevor, R. Tibshirani, and J. Friedman. 2001. *The Elements of Statistical Learning*. New York: Springer Verlag.

Heckman, James, H. Ichimura, and P. Todd. 1998. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program." *Review of Economic Studies* 64:605–654.

Heckman, James J., Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrika* 66(5):1017–1098.

Ho, Daniel, Kosuke Imai, Gary King, and Elizabeth Stuart. 2005. "Matching as Nonparametric Preprocessing for Parametric Causal Inference." http://gking.harvard.edu/files/matchp.pdf.

Hoeting, Jennifer A., David Madigan, Adrian E. Raftery, and Chris T. Volinsky. 1999. "Bayesian Model Averaging: A Tutorial." *Statistical Science* 14(4):382–417.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81:945–960.

Imai, Kosuke, and David A. van Dyk. 2004. "Causal Inference with General Treatment Regimes: Generalizing the Propensity Score." *Journal of the American Statistical Association* 99(467):854–866.

Imai, Kosuke, and Gary King. 2004. "Did Illegal Overseas Absentee Ballots Decide the 2000 U.S. Presidential Election?" *Perspectives on Politics* 2(3):537–549.

Kallay, Michael. 1986. "Convex Hull Made Easy." *Information Processing Letters* 22(March):161.

King, Gary. 1991. "'Truth' Is Stranger than Prediction, More Questionable than Causal Inference." *American Journal of Political Science* 35(4):1047–1053. http://gking.harvard.edu/files/abs/truth-abs.shtml.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton, NJ: Princeton University Press.

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44(2):341–355. http://gking.harvard.edu/files/abs/making-abs.shtml.

King, Gary, and Langche Zeng. 2002. "Improving Forecasts of State Failure." *World Politics* 53(4):623–658. http://gking.harvard.edu/files/abs/civil-abs.shtml.

Klee, Victor. 1980. "On the Complexity of d-Dimensional Voronoi Diagrams." *Archive der Mathematik* 34:75–80.

Kuo, Yen-Hong. 2001. "Extrapolation of Association between Two Variables in Four General Medical Journals." Presented at the Fourth International Congress on Peer Review in Biomedical Publication, Barcelona, Spain.

Lechner, Michael. 1999. "Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumptions." IZA Discussion Papers no. 91, University St. Gallen.

Madych, W. R., and S. A. Nelson. 1992. "Bounds on Multivariate Polynomials and Exponential Error Estimates for Multiquadric Interpolation." *Journal of Approximation Theory* 70:94–114.

Manski, Charles F. 1995. *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.

Meng, Xiao-Li, and Marin Romero. 2003. "Discussion: Efficiency and Self-Efficiency." *International Statistical Review* 71(3):607–618.

O'Rourke, Joseph. 1998. *Computational Geometry in C*. New York: Cambridge University Press.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge, UK: Cambridge University Press.

Robins, James M. 1999a. "Marginal Structural Models versus Structural Nested Models as Tools for Causal Inference." In *Statistical Models in Epidemiology: The Environment and Clinical Trials*, vol. 16, eds. M. E. Halloran and D. Berry. New York: Springer-Verlag, pp. 95–134.

Robins, James M. 1999b. "Association, Causation, and Marginal Structural Models." *Synthese* 121:151–179.

Rosenbaum, Paul. 1984. "The Consequences of Adjusting for a Concomitant Variable That Has Been Affected by the Treatment." *Journal of the Royal Statistical Society, A* 147(5):656–666.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70:41–55.

Rosenbaum, Paul R., and Donald B. Rubin. 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association* 79:515–524.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 6:688–701.

Schaback, R. 1996. "Approximation by Radia Basis Functions with Finitely Many Centers." *Constructive Approximation* 12:331–340.

Sisson, Scott A. 2005. "Transdimensional Markov Chains: A Decade of Progress and Future Perspectives." *Journal of the American Statistical Association* 100(471):1077–1089.

Stoll, Heather, Gary King, and Langche Zeng. 2005. "WhatIf: Software for Evaluating Counterfactuals." http://gking.harvard.edu/whatif/.

Valentine, Frederick Albert. 1964. *Convex Sets*. New York: McGraw-Hill.

Winship, Christopher, and Stephen L. Morgan. 1999. "The Estimation of Causal Effects from Observational Data." *American Review of Sociology* 25:659–707.

Wu, Z., and R. Schaback. 1993. "Local Error Estimates for Radial Basis Function Interpolation of Scattered Data." *Journal of Numerical Analysis* 13:13–27.